ELSEVIER

Signal Processing 80 (2000) 2219–2235

SIGNAL PROCESSING

www.elsevier.nl/locate/sigpro

Coefficient of determination in nonlinear signal processing

Edward R. Dougherty[a,*], Seungchan Kim[a], Yidong Chen[b]

[a]*Department of Electrical Engineering, Texas A&M University, College Station, TX 77843-3128, USA*
[b]*National Human Genome Research Institute, National Institutes of Health, USA*

Received 16 August 1999; received in revised form 14 February 2000

Abstract

For statistical design of an optimal filter, it is probabilistically advantageous to employ a large number of observation random variables; however, estimation error increases with the number of variables, so that variables not contributing to the determination of the target variable can have a detrimental effect. In linear filtering, determination involves the correlation coefficients among the input and target variables. This paper discusses use of the more general coefficient of determination in nonlinear filtering. The determination coefficient is defined in accordance with the degree to which a filter estimates a target variable beyond the degree to which the target variable is estimated by its mean. Filter constraint decreases the coefficient, but it also decreases estimation error in filter design. Because situations in which the sample is relatively small in comparison with the number of observation variables are of salient interest, estimation of the determination coefficient is considered in detail. One may be unable to obtain a good estimate of an optimal filter, but can nonetheless use rough estimates of the coefficient to find useful sets of observation variables. Since minimal-error estimation underlies determination, this material is at the interface of signal processing, computational learning, and pattern recognition. Several signal-processing factors impact application: the signal model, morphological operator representation, and desirable operator properties. In particular, the paper addresses the VC dimension of increasing operators in terms of their morphological kernel/basis representations. Two applications are considered: window size for restoring degraded binary images; finding sets of genes that have significant predictive capability relative to target genes in genomic regulation.

Zusammenfassung

Für den statistischen Entwurf eines optimalen Filters ist es im probabilistischen Sinn vorteilhaft, eine große Anzahl von Beobachtungs-Zufallsvariablen zu verwenden. Der Schätzfehler steigt jedoch mit der Anzahl der Variablen, so daß Variablen, die nicht zur Bestimmung der Zielvariablen beitragen, einen nachteiligen Effekt haben können. Bei der linearen Filterung involviert die Bestimmung die Korrelationskoeffizienten der Eingangs- und Zielvariablen. In diesem Artikel wird die Verwendung des allgemeineren Bestimmungskoeffizienten für die nichtlineare Filterung diskutiert. Der Bestimmungskoeffizient ist definiert gemäß dem Grad, mit welchem die Zielvariable durch ein Filter besser geschätzt wird als durch den Mittelwert der Zielvariablen. Eine Einschränkung des Filters verringert diesen Koeffizienten, gleichzeitig aber auch den Schätzfehler beim Filterentwurf. Da Situationen, in denen die Stichprobe im Vergleich zur Anzahl der Beobachtungsvariablen klein ist, von besonderem Interesse sind, wird die Schätzung des Bestimmungskoeffizienten im Detail betrachtet. Es ist möglich, daß man zwar keinen guten Schätzer eines Optimalfilters erhalten

* Corresponding author. Tel.: 1-409-862-8154; fax: +1-409-862-3336.
E-mail address: e-daughesty@tanu.edu (E.R. Dougherty).

0165-1684/00/$ - see front matter
PII: S0165-1684(00)00079-7

kann, jedoch trotzdem grobe Schätzwerte des Koeffizienten verwenden kann, um nützliche Mengen von Beobachtungsvariablen zu finden. Da die Schätzung mit minimalem Fehler der Bestimmung zugrundeliegt, befindet sich diese Materie an der Schnittstelle von Signalverarbeitung, automatisiertem Lernen und Mustererkennung. Mehrere Signalverarbeitungsfaktoren wirken sich auf die Anwendung aus: Signalmodell, morphologische Operatordarstellung und erwünschte Operatoreigenschaften. Der Artikel behandelt insbesondere die VC-Dimension wachsender Operatoren unter Verwendung ihrer morphologischen Kern- und Basisdarstellungen. Zwei Anwendungen werden betrachtet: die Fenstergröße bei der Rekonstruktion gestörter binärer Bilder sowie in der Genom-Regulierung die Ermittlung von Gensätzen, die signifikante prädiktive Eigenschaften bezüglich Zielgenen besitzen.

Résumé

Pour de la conception statistique de filtre optimal, il est probabilistiquement avantageux d'employer un large nombre de variables d'observations aléatoires; cependant, l'erreur d'estimation croît avec le nombre de variables, de sorte que les variables qui ne contribuent pas à la détermination de la variable cible peuvent avoir un effet négatif. En filtrage linéaire, cette détermination implique les coefficients de corrélation entre les variables d'entrée et cibles. Cet article traite de l'utilisation du coefficient plus général de détermination en filtrage non-linéaire. Le coefficient de détermination est défini en accord avec le degré auquel le filtre estime une variable cible au-delà du degré auquel la variable cible est estimée par sa moyenne. La contrainte du filtre diminue le coefficient, mais aussi l'erreur d'estimation en conception de filtres. Parce que les situations ou l'échantillon est relativement petit en comparaison du nombre de variables d'observations sont d'un intérêt majeur, l'estimation du coefficient de détermination est considérée en détail. On peut être incapable d'obtenir une bonne estimation du filtre optimal mais on peut cependant utiliser une estimation grossière du coefficient pour trouver des ensembles utiles de variables d'observation. Puisque l'estimation à erreur minimale sous-tend la détermination, ce matériel est à l'interface du traitement des signaux, de l'apprentissage informatique et de la reconnaissance des formes. Plusieurs facteurs de traitement de signaux ont un impact sur l'application: le modèle de signal, la représentation d'opérateurs morphologiques, et les propriétés désirables des opérateurs. En particulier, cet article traite de la dimension VC d'opérateurs croissants en terme de lcurs représentations en noyaux/bases morphologiques. Deux applications sont considérées: la taille de la fenêtre pour restaurer des images binaires et la recherche d'ensembles de gènes qui ont une capacité prédictive significative relative à des gènes cibles en régulation génomique.

Keywords: Estimation; Genomics; Mathematical morphology; Optimal filter; VC dimension

1. Introduction

A fundamental problem of nonlinear digital signal (image) processing is the automatic design of filters to estimate an ideal random signal F from an observed random signal G. Significant effort has gone into designing window-based filters. For these, an n-point window W is placed at a point z, thereby determining a random vector $\boldsymbol{X}$ of G-values in the window. A *computational operator* ψ is applied to $\boldsymbol{X}$ to form an estimator $\psi(\boldsymbol{X})$ of the value $Y = F(z)$. The *W-operator* Ψ is defined by $\Psi(G)(z) = \psi(\boldsymbol{X})$. If F and G are jointly stationary, then ψ is independent of z. The difficulty is that filter design usually depends on estimating a large number of parameters. The number grows exponentially with the number of observation random variables composing $\boldsymbol{X}$. The problem is mitigated by optimizing over constrained classes of filters and by employing prior information; nonetheless, a substantial amount of data is often required to obtain precise (close-to-optimal) designed filters.

One would like to use as large a window as possible; however, as the window size grows, so does the estimation error (for fixed sample size). Thus, when considering window enlargement, it is beneficial to only adjoin points whose values provide more than a negligible increase in the determination of the target value in the ideal image. This paper discusses the coefficient of determination in the context of nonlinear digital signal and image processing.

Filter design is a form of inverse problem. We consider F to be operated on by a random *system transformation* Ξ to produce the observed signal $G = \Xi F$. The inverse problem is to find an optimal estimator $\Psi(G)$ for F. For a W-operator, the joint random signal process (F, G) induces a distribution on the $(n+1)$-vector $(\boldsymbol{X}, Y)$. This induced distribution determines the optimal filter. Often, we know nothing about the distribution; however, in many cases there are signal properties which imply that the optimal filter belongs to some subclass C of filters. For instance, Ξ might be such that the optimal filter must be increasing. If so, then we can estimate the optimal increasing filter and thereby lessen the amount of sample data required for precise estimation, without sacrificing the goal of optimality. More generally, from either theory or experience, we might know that C will provide (or likely provide) a good suboptimal filter.

Recognizing the role of the signals in the filter-design paradigm is critical for appreciating estimation of the coefficient of determination as opposed to estimation of an optimal filter. For digital signal processing, the logical structure of an optimal filter reflects the structural relations between F and G. This is perhaps best appreciated in morphological image processing, where Ψ is defined by structuring elements (geometric templates) [48]. These structuring elements exhibit completely the manner in which the structure of G is to be transformed to estimate the structure of F. Equivalently, they reflect the manner in which the structural deformation of Ξ can be best inverted. The error of the optimal filter reflects the degree to which the structural deformation can be inverted by logical operations upon the random variables in the window. A larger window is beneficial because it provides greater structural transformation, but statistical estimation of beneficial structural transformations becomes rapidly more difficult for increasing window size.

For samples that are too small relative to the number of observation variables, it is not possible to obtain a good estimate of an optimal filter; however, it still may be possible to decide which observations contribute more to the determination of the unobserved variable. This information alone will not tell us how to transform the observation templates, but it will tell us the degree to which they can be transformed to restore the ideal signal. For structural properties, the size and shape of the window are critical, and even if we only have crude estimates of the coefficients of determination for various windows, we may still have enough information to decide the relative benefit of different windows. Perhaps this is most evident in binary image processing, where, for a fixed window size, different window shapes may reveal to very different degrees the relationship between the geometries of the observed and unobserved images.

Because this paper is at the interface between nonlinear signal processing, computational learning, and pattern recognition (in particular, the Vapnik–Chervonenkis theory), it is important to recognize the role of constraint in the statistical design of nonlinear filters. Most relevant is the manner in which filter constraints occur naturally in accordance with image structure, algebraic properties of operators, and morphological operator representation. For the general case of translation-invariant binary operators, morphological representation restricted to finite windows becomes an extension of Boolean representation [2]. If one only considers $\boldsymbol{X}$ as a binary vector and Y as a binary random variable to be estimated by a trained operator, then statistical design lies in the domain of computational learning, albeit, with the representational structure of mathematical morphology [5,6,13,16,27]. A similar statement applies to the decomposition of gray-scale operators and their unconstrained window design [3,14,19]. But for the most part, this abstract perspective is not germane in practice owing to the quantity of sample data necessary for precise filter estimation [17]. Various techniques are applied to mitigate the data demand and enhance the outputs: structural constraints on the operators [38], prior knowledge concerning filter structure [4,18], and partial constraint for some observation variables [46]. These constraints evolve out of signal processing considerations, including desirable operator properties.

The increasing constraint has played a key role in image operator theory, and morphological erosion representation of increasing operators pre-dates the general representation of arbitrary

translation-invariant operators, first in terms of the operator kernel [43], and then in terms of its basis [24,41,42]. Owing to their prevalence and early morphological representation, automatic design of increasing operators came first [7,10,11]. Binary increasing filters, and stack filters, which are essentially binary because a single Boolean function operates on threshold sets, have been especially studied owing to their geometrical nature, smaller data requirements, and less complex architecture [1,7]. Structural constraints have played a key role [23,34–36,52]. Design tools have been developed to determine acceptable suboptimal estimates of an optimal increasing filter: recursive error estimation, [37] iterative decomposition [21], adaptive design [44], and genetic algorithms [26,33]. Direct filter synthesis from an image model has also been considered [12]. There are image models in which the optimal increasing filter is fully optimal, and others where it is close to optimal, and therefore it is a natural constraint. In addition, logical implementation of satisfactory suboptimal increasing filters is often extremely less complex than implementation of corresponding non-increasing filters. While the increasing constraint occurs naturally in the context of signal and image processing, the constraint is also naturally viewed in the context of pattern recognition, and to that extent we will consider the VC dimension of increasing filters in terms of their morphological erosion bases.

For an instance from natural science where measuring determination can be beneficial, we consider genomics, where the mRNA expression levels of the full genome can be treated as a random time vector [15]. One problem is to predict the expression level Y of a specific gene at time t_1 from the expression levels of a vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ of expression levels at time $t_0 < t_1$. One can also make predictions across the genome at a given time. While geneticists would like to have logical models by which one gene expression is predicted by a set of expressions, both technology and cost mitigate against the size of experiments necessary for precise estimation of optimal predictors (filters) – for instance, with cDNA microarrays [8,22,47]. Nonetheless, insight into the regulatory mechanisms of the genome can be gained when there is knowledge as to which gene sets are regulatory with respect to specific genes, especially when this is combined with biological knowledge. Moreover, appreciation of the relative degrees of determination between different predictor gene sets can help to design future experimentation via the formulation of working hypotheses within the context of functional genomics.

2. Coefficient of determination

To define a measure of determination, let Y be a random variable to be estimated via a subset $\mathscr{X}$ of a family $\mathscr{V}$ of conditioning (observation) random variables. A filter (estimator) $\psi(\mathscr{X})$ is formed by a (measurable) function ψ whose domain is the product space of the ranges of $\mathscr{X}$, and whose range is the range of Y. The goodness of $\psi(\mathscr{X})$ is typically quantified by an estimation error that is the expectation of a loss function $l(a, b)$ measuring the cost of the difference between a and b: $\varepsilon[Y,\psi(\mathscr{X})] = E[l(Y, \psi(\mathscr{X}))]$. $\varepsilon[\psi]$ denotes the error if Y and $\mathscr{X}$ are clear from the context. Examples are the mean-square error (MSE) $E[|Y - \psi(\mathscr{X})|^2]$ and mean-absolute error (MAE) $E[|Y - \psi(\mathscr{X})|]$, which we denote by $M[\psi]$. Unconstrained optimization results from allowing ψ to be any measurable function of the random variables of $\mathscr{X}$ and choosing a function $\psi_{\mathscr{X}}$ having minimal error. For unconstrained optimization, there is a basic monotonicity property: if $\mathscr{X} \subset \mathscr{Z} \subset \mathscr{V}$, then $\varepsilon[\psi_{\mathscr{Z}}] \leqslant \varepsilon[\psi_{\mathscr{X}}]$. This property applies to $\mathscr{X} = \emptyset$, in which case, $\psi_\emptyset$ is a constant that minimizes $\varepsilon[Y, c]$ over all constants c. Often ψ is constrained to a subclass of the class of all functions of the random variables. For each $\mathscr{X} \subset \mathscr{V}$, there is a function class $C(\mathscr{X})$ from which ψ is chosen. $\psi_{C(\mathscr{X})}$ denotes an optimal filter from $C(\mathscr{X})$. Mathematically, a constraint C is a function whose domain is the set of subsets of $\mathscr{V}$ and, for each $\mathscr{X} \subset \mathscr{V}$, $C(\mathscr{X})$ is a subset of the set of all functions of $\mathscr{X}$. A constraint is *nested* if $\mathscr{X} \subset \mathscr{Z} \subset \mathscr{V}$ implies $C(\mathscr{X}) \subset C(\mathscr{Z})$. If C is nested, then a monotonicity property applies: $\varepsilon[\psi_{C(\mathscr{Z})}] \leqslant \varepsilon[\psi_{C(\mathscr{X})}]$ for $\mathscr{X} \subset \mathscr{Z}$. We assume nested constraints. Unconstrained optimization is a special case of constrained optimization with $C(\mathscr{X})$ being the class of all functions on $\mathscr{X}$, for all $\mathscr{X} \subset \mathscr{V}$. If we wish to emphasize the constraint and the conditioning set,

and not the optimal filter itself, then we may write $\varepsilon[C, \mathscr{X}]$ for the error of the optimal filter $\psi_{C(\mathscr{X})}$.

Two constrained function classes $C(\mathscr{X})$ and $D(\mathscr{X})$ can be partially ordered by the subset relation, meaning that $C(\mathscr{X}) \subset D(\mathscr{X})$. In this case, for fixed $\mathscr{X}$, $\varepsilon[\psi_{D(\mathscr{X})}] \leqslant \varepsilon[\psi_{C(\mathscr{X})}]$. This partial order extends to the constraints themselves: $C \leqslant D$ if $C(\mathscr{X}) \subset D(\mathscr{X})$ for all $\mathscr{X} \subset \mathscr{V}$.

Consider a nested constraint C in which the constant function class $C(\emptyset)$ is unconstrained. If the constant λ_Y denotes the best constant estimate of Y, then the *coefficient of determination* of Y relative to the conditioning set $\mathscr{X}$ for the constraint C is defined by

$$\theta[C,\mathscr{X}] = \frac{\varepsilon[\lambda_Y] - \varepsilon[\psi_{C(\mathscr{X})}]}{\varepsilon[\lambda_Y]} = \frac{\varepsilon[C, \emptyset] - \varepsilon[C,\mathscr{X}]}{\varepsilon[C, \emptyset]}. \tag{1}$$

It measures the relative decrease in error from estimating Y via $\mathscr{X}$ by $\psi_{C(\mathscr{X})}$, rather than just by λ_Y. Owing to nestedness, $0 \leqslant \theta[C, \mathscr{X}] \leqslant 1$. Specifically, if $\varepsilon[\psi_{C(\mathscr{X})}] = 0$, then $\theta[C, \mathscr{X}] = 1$. If C and $\mathscr{X}$ are contextually clear, we just write θ. For fixed $\mathscr{X}$ and constraints C and D, if $C(\mathscr{X}) \subset D(\mathscr{X})$, then $\theta[C, \mathscr{X}] \leqslant \theta[D, \mathscr{X}]$. Because C is nested, if $\mathscr{X} \subset \mathscr{Z}$, then $\theta[C, \mathscr{X}] \leqslant \theta[C, \mathscr{Z}]$. Thus, for constraint C, we can define the *incremental determination* for $\mathscr{Z}$ relative to $\mathscr{X}$ by

$$\theta^+[C, \mathscr{X}, \mathscr{Z}] = \theta[C, \mathscr{Z}] - \theta[C, \mathscr{X}]. \tag{2}$$

Note that, $\theta^+[C, \emptyset, \mathscr{Z}] = \theta[C, \mathscr{Z}]$.

For real-valued random variables, MSE, and an unbiased optimal estimator, determination can be represented in terms of variances. Since $\psi_{C(\mathscr{X})}$ is unbiased, $E[\psi_{C(\mathscr{X})}(\mathscr{X})] = \mu_Y$, the mean of Y, and $E[Y - \psi_{C(\mathscr{X})}(\mathscr{X})] = 0$. Thus,

$$\begin{aligned}\varepsilon[\psi_{C(\mathscr{X})}] &= E[|Y - \psi_{C(\mathscr{X})}(\mathscr{X})|^2] \\ &= \mathrm{Var}[Y - \psi_{C(\mathscr{X})}(\mathscr{X})].\end{aligned} \tag{3}$$

Moreover, $\lambda_Y = \mu_Y$ and $\varepsilon[\lambda_Y] = \varepsilon[\mu_Y] = \sigma_Y^2 = \mathrm{Var}[Y]$. Thus,

$$\theta[C, \mathscr{X}] = \frac{\mathrm{Var}[Y] - \mathrm{Var}[Y - \psi_{C(\mathscr{X})}(\mathscr{X})]}{\mathrm{Var}[Y]}. \tag{4}$$

If $\psi_{C(\mathscr{X})}(\mathscr{X})$ and $Y - \psi_{C(\mathscr{X})}(\mathscr{X})$ are uncorrelated, then the numerator reduces to $\mathrm{Var}[\psi_{C(\mathscr{X})}(\mathscr{X})]$ and

$$\theta[C,\mathscr{X}] = \frac{\mathrm{Var}[\psi_{C(\mathscr{X})}(\mathscr{X})]}{\mathrm{Var}[Y]} \tag{5}$$

which is the proportion of $\mathrm{Var}[Y]$ "explained" by the optimal filter. For a single observation X for which X and Y are jointly Gaussian, θ is the square of the correlation coefficient for X and Y. It is via Eq. (4) that the coefficient of determination has been used to measure the significance of multiple linear regression [51]. No such simple reduction occurs for nonlinear digital signal filters.

3. Estimation error: unconstrained filters

A designed estimate of the optimal filter is derived from sample data. For an observation vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, estimation is typically achieved by applying an estimation rule to a random sample S of vector-variable pairs identically distributed to $(\boldsymbol{X}, Y)$ to estimate the parameters determining the optimal filter. The goodness of the estimation depends on the sample size. For a sample size of N, we obtain an estimate ψ_n^N of the optimal filter, ψ_n, and the error of the estimated filter is decomposed as

$$\varepsilon[\psi_n^N] = \varepsilon[\psi_n] + \Delta(\psi_n^N,\psi_n), \tag{6}$$

where $\Delta(\psi_n^N,\psi_n)$ is the estimation error. $\Delta(\psi_n^N,\psi_n)$ depends on the estimation procedure for ψ_n. Since ψ_n^N depends on S, so does $\Delta(\psi_n^N,\psi_n)$. We consider the expectation $E[\Delta(\psi_n^N,\psi_n)]$ of $\Delta(\psi_n^N,\psi_n)$ over all samples. We could write $\psi_n^{N,S}$, $\varepsilon[\psi_n^{N,S}]$, $\Delta(\psi_n^{N,S},\psi_n)$, and $E_S[\Delta(\psi_n^{N,S},\psi_n)]$ to indicate the role of the sample S; however, to ease notation we leave S implicit in the notation.

Using Eq. (6), the coefficient of determination is estimated by

$$\begin{aligned}\theta_n^N &= \frac{\varepsilon[\psi_0^N] - \varepsilon[\psi_n^N]}{\varepsilon[\psi_0^N]} \\ &= \frac{\varepsilon[\psi_0] - \varepsilon[\psi_n] + \Delta(\psi_0^N,\psi_0) - \Delta(\psi_n^N,\psi_n)}{\varepsilon[\psi_0] + \Delta(\psi_0^N,\psi_0)}.\end{aligned} \tag{7}$$

Since this estimator depends on the sample, it is a random variable (as are the estimation errors). Thus, our concern is with the expectation $E[\theta_n^N]$.

Owing to the random variable $\Delta(\psi_0^N,\psi_0)$ in the denominator of θ_n^N, analysis of $E[\theta_n^N]$ is problematic. However, usually there are very few parameters in ψ_0^N, and ψ_0^N can be precisely estimated. In this case, we can obtain a good approximation by letting $\Delta(\psi_0^N,\psi_0)=0$; namely,

$$\theta_n^N \approx \frac{\varepsilon[\psi_0]-\varepsilon[\psi_n]-\Delta(\psi_n^N,\psi_n)}{\varepsilon[\psi_0]}, \tag{8}$$

Taking expectations yields

$$E[\theta_n^N] \approx \theta_n - \frac{E[\Delta(\psi_n^N,\psi_n)]}{\varepsilon[\psi_0]}. \tag{9}$$

Since $E[\Delta(\psi_n^N,\psi_n)] > 0$, θ_n^N is biased low as an estimator of θ_n, the bias being the quotient in Eq. (9). For consistent estimation rules, $E[\Delta(\psi_n^N,\psi_n)] \to 0$ as $N\to\infty$, and therefore θ_n^N is asymptotically unbiased. For small samples the bias can be significant. Hence, the question arises as to how fast $E[\theta_n^N]\to\theta_n$, or, equivalently, how fast $\varepsilon[\psi_n^N]\to\varepsilon[\psi_n]$. We will continue to assume the suitability of the approximation $\Delta(\psi_0^N,\psi_0)=0$.

Without distributional assumptions on $(\boldsymbol{X}, Y)$, convergence can be very slow. To see this, suppose $\boldsymbol{X}$ has real-valued components, Y is binary, and error is MAE. Let κ_n^N be an estimator of $M[\psi_n]$ based on a random sample of size N. Assuming consistency, $\kappa_n^N - M[\psi_n]\to 0$ as $N\to\infty$. The following theorem shows that convergence is arbitrarily slow [9]: for any estimator κ_n^N and for every $\rho>0$, there exists a distribution of $(\boldsymbol{X}, Y)$ such that

$$E[|\kappa_n^N - M[\psi_n]|] \geqslant \tfrac{1}{4}-\rho. \tag{10}$$

This bound applies at once to Eq. (9) to yield

$$\theta_n - E[\theta_n^N] \geqslant \frac{1-4\rho}{4M[\psi_0]}. \tag{11}$$

4. Coefficient of determination for unconstrained binary filters

For a binary random vector $\boldsymbol{X}=(X_1,X_2,\ldots,X_n)$ and a binary random variable Y, the optimal unconstrained MAE filter is the binary conditional expectation: $\psi_n(\boldsymbol{x})=1$ if $P(Y=1|\boldsymbol{x})>0.5$ and $\psi_n(\boldsymbol{x})=0$ if $P(Y=1|\boldsymbol{x})\leqslant 0.5$. Implicit in the formulation is that, for the null vector, $\boldsymbol{X}^{(0)}$, $\psi_0(\boldsymbol{X}^{(0)})$ is the thresholded mean of Y. For any binary-valued operator ψ, its kernel is defined by $\mathscr{K}[\psi]=\{\boldsymbol{x}\colon \psi(\boldsymbol{x})=1\}$. For the optimal filter, $\mathscr{K}[\psi_n]=\{\boldsymbol{x}\colon P(Y=1|\boldsymbol{x})>0.5\}$.

For $n>0$, the MAE for ψ_n is expressed in terms of the kernel by

$$M[\psi_n] = \sum_{\boldsymbol{x}\in\mathscr{K}[\psi_n]} P(\boldsymbol{x})P(Y=0|\boldsymbol{x}) + \sum_{\boldsymbol{x}\notin\mathscr{K}[\psi_n]} P(\boldsymbol{x})P(Y=1|\boldsymbol{x}). \tag{12}$$

It can be equivalently expressed as

$$M[\psi_n] = \sum_{\boldsymbol{x}} P(\boldsymbol{x})\min\{P(Y=0|\boldsymbol{x}), P(Y=1|\boldsymbol{x})\}. \tag{13}$$

This formulation is consistent with the MAE for the null case, since $\mu_Y = P(Y=1)$ and

$$M[\psi_0] = M[\lambda_Y] = \min\{P(Y=0), P(Y=1)\}. \tag{14}$$

Using the fact that the total probability is one, the determination can be expressed as

$$\theta_n = \frac{\sum_{\boldsymbol{x}} P(\boldsymbol{x})[\min\{P(Y=0), P(Y=1)\} - \min\{P(Y=0|\boldsymbol{x}), P(Y=1|\boldsymbol{x})\}]}{\min\{P(Y=0), P(Y=1)\}}, \tag{15}$$

where we assume that $0<\mu_Y<1$, so that $\min\{P(Y=0), P(Y=1)\}>0$.

For any $\boldsymbol{x}$, the term of the sum corresponding to $\boldsymbol{x}$ is the *contribution* of $\boldsymbol{x}$ to θ_n. We denote it by $\theta_n(\boldsymbol{x})$. Suppose $\mu_Y\geqslant 0.5$ and $\boldsymbol{x}\in\mathscr{K}[\psi_n]$. Then the first and second minimums in the numerator of Eq. (15) are $P(Y=0)$ and $P(Y=0|\boldsymbol{x})$, respectively, and

$$\theta_n(\boldsymbol{x}) = \frac{P(Y=1|\boldsymbol{x})-P(Y=1)}{\mu_Y}P(\boldsymbol{x}). \tag{16}$$

If $P(Y=1|\boldsymbol{x})>P(Y=1)$, then $\theta_n(\boldsymbol{x})>0$. Conditioning by $\boldsymbol{x}$ increases the extent to which the probability of Y equaling 1 exceeds 0.5, and this is

reflected in a positive contribution for $\boldsymbol{x}$. If, on the other hand, $P(Y=1|\boldsymbol{x}) < P(Y=1)$, then $\theta_n(\boldsymbol{x}) < 0$. Here, conditioning by $\boldsymbol{x}$ decreases the extent to which the probability of Y equaling 1 exceeds 0.5. The situations for $\mu_Y \geqslant 0.5$ and $\boldsymbol{x} \notin \mathcal{K}[\psi_n]$,$\mu_Y \leqslant 0.5$ and $\boldsymbol{x} \in \mathcal{K}[\psi_n]$, and $\mu_Y \leqslant 0.5$ and $\boldsymbol{x} \notin \mathcal{K}[\psi_n]$ can be similarly analyzed.

For optimal unconstrained binary filtering, we need to estimate the conditional probabilities $P(Y=1|\boldsymbol{x})$ and the observation probabilities $P(\boldsymbol{x})$. The estimation error is

$$\Delta(\psi_n^N, \psi_n) = \sum_{\boldsymbol{x} \in \mathcal{K}[\psi_n] \underline{\Delta} \mathcal{K}[\psi_n^N]} |1 - 2P(Y=1|\boldsymbol{x})| P(\boldsymbol{x}), \quad (17)$$

where $\underline{\Delta}$ denotes the symmetric difference between the kernels. The error depends on how ψ_n^N is estimated from the sample S. This can be done using the sample probability estimates $P_S(\boldsymbol{x})$ and $P_S(Y=1|\boldsymbol{x})$ computed from S. Then ψ_n^N is determined by using $P_S(Y=1|\boldsymbol{x})$ in place of $P(Y=1|\boldsymbol{x})$ in the definition of ψ_n. The problem with this approach is that, for even modestly large windows, the training set will be too small to obtain good estimates; in fact, there will be many vectors never observed in training. In the context of pattern recognition, there are various estimation rules for $P(Y=1|\boldsymbol{x})$ to circumvent this problem for the conditional probabilities. In signal processing, when there is insufficient data to estimate $P(Y=1|\boldsymbol{x})$ for a vector $\boldsymbol{x}$, a number of methods can be used to define $\psi_n^N(\boldsymbol{x})$, including the use of prior signal information.

A case in point is differencing-filter design, in which $\psi_n^N(\boldsymbol{x})$ is defined to be the component value x_0 of $\boldsymbol{x}$ corresponding to the window center when the estimate $P_S(Y=1|\boldsymbol{x})$ lacks sufficient credibility. This means the pixel value is passed unless there is sufficient reason to change it. Differencing filters have proven useful for the restoration and resolution conversion of digital documents [40]. If the criterion to use $P_S(Y=1|\boldsymbol{x})$ to define $\psi_n^N(\boldsymbol{x})$ is that $N_{\boldsymbol{x}}$, the number of times $\boldsymbol{x}$ is observed during training, is at least τ, then taking the expectation in Eq. (17) yields

$$\begin{aligned} E[\Delta(\psi_n^N, \psi_n)] &= \sum_{\{\boldsymbol{x}: P(Y=1|\boldsymbol{x}) > 0.5, x_0 = 1\}} \delta_{\boldsymbol{x}} P(P_S(Y=1|\boldsymbol{x}) \leqslant 0.5, N_{\boldsymbol{x}} \geqslant \tau) \\ &+ \sum_{\{\boldsymbol{x}: P(Y=1|\boldsymbol{x}) \leqslant 0.5, x_0 = 0\}} \delta_{\boldsymbol{x}} P(P_S(Y=1|\boldsymbol{x}) > 0.5, N_{\boldsymbol{x}} \geqslant \tau) \\ &+ \sum_{\{\boldsymbol{x}: P(Y=1|\boldsymbol{x}) > 0.5, x_0 = 0\}} \delta_{\boldsymbol{x}} [P(N_{\boldsymbol{x}} < \tau) + P(P_S(Y=1|\boldsymbol{x}) \leqslant 0.5, N_{\boldsymbol{x}} \geqslant \tau)] \\ &+ \sum_{\{\boldsymbol{x}: P(Y=1|\boldsymbol{x}) \leqslant 0.5, x_0 = 1\}} \delta_{\boldsymbol{x}} [P(N_{\boldsymbol{x}} < \tau) + P(P_S(Y=1|\boldsymbol{x}) > 0.5, N_{\boldsymbol{x}} \geqslant \tau)], \end{aligned} \quad (18)$$

where $\delta_{\boldsymbol{x}} = |2P(Y=1|\boldsymbol{x}) - 1|$ [18]. Differencing design yields consistent estimation of $\varepsilon[\psi_n]$.

Returning to the general case, it is always true that $0 \leqslant \theta_1 \leqslant \theta_2 \leqslant \ldots \leqslant \theta_n \leqslant 1$; however, this increasing determination relative to the number of observations can be problematic when using sample data. According to Eq. (9), the difference between $E[\theta_k^N]$ and $E[\theta_{k+1}^N]$ is given by

$$E[\theta_{k+1}^N] - E[\theta_k^N] \approx \theta_{k+1} - \theta_k - \frac{E[\Delta(\psi_{k+1}^N, \psi_{k+1})] - E[\Delta(\psi_k^N, \psi_k)]}{\varepsilon[\psi_0]}. \quad (19)$$

The preceding numerator is nonnegative, and very likely positive. If $\theta_{k+1} - \theta_k$ is very small, and $\Delta(\psi_0^N, \psi_0) \approx 0$ (so that the approximation is accurate), the difference between the expectations can be negative, thereby yielding expected determination coefficients that are not monotone. Using more variables is only beneficial, relative to $E[\theta_n^N]$, if the increased determination more than offsets the increased estimation error.

5. Estimation error: constrained filters

Let $\psi_{n,C}$ denote the optimal filter in a subclass C. There is a cost of constraint, $\Delta(\psi_{n,C}, \psi_n)$, and

$$\varepsilon[\psi_{n,C}] = \varepsilon[\psi_n] + \Delta(\psi_{n,C}, \psi_n). \quad (20)$$

There is an estimation cost $\Delta(\psi_{C,n}^N, \psi_{n,C})$ for $\psi_{n,C}$. An analogue of Eq. (6) applies. Moreover,

$$\varepsilon[\psi_{C,n}^N] = \varepsilon[\psi_n] + \Delta(\psi_{n,C}, \psi_n) + \Delta(\psi_{n,C}^N, \psi_{n,C}). \quad (21)$$

Hence, the expected estimate of the determination coefficient if we use a filter from C is

$$E[\theta_{n,C}^N] \approx \theta_n - \frac{\Delta(\psi_{n,C}, \psi_n) + E[\Delta(\psi_{n,C}^N, \psi_{n,C})]}{\varepsilon[\psi_0]}. \quad (22)$$

The constraint is beneficial if and only if

$$\Delta(\psi_{n,C},\psi_n) + E[\Delta(\psi_{n,C}^N,\psi_{n,C})] < E[\Delta(\psi_n^N,\psi_n)]. \quad (23)$$

If the optimal constrained filter is not much worse than the optimal filter and there is sufficient improvement in estimation error, constraint is beneficial. In many situations, the estimation error is so great that constraint is necessary.

Insofar as estimation error for constrained filters is involved, the Vapnik–Chervonenkis theory applies. Consider MAE and, for $\psi \in C$, define the *empirical error estimate* for ψ to be $1/N$ times the number of errors made by ψ on the sample data. Suppose the designed filter ψ_n^N is chosen as the one having minimal empirical error (which was done when we used the sample probability estimates for binary-filter design). The *Vapnik–Chervonenkis theorem* states that, for $\rho > 0$,

$$P(\Delta(\psi_{n,C}^N, \psi_{n,C}) > \rho) \leqslant 8S(C,N)\exp\left[-\frac{N\rho^2}{32}\right], \quad (24)$$

where $S(C,N)$ is the *shatter coefficient* of the class C (to be explained shortly) [50]. The theorem bounds the probability that $\Delta(\psi_{n,C}^N,\psi_{n,C})$ exceeds ρ in terms of the shatter coefficient and an exponential of $-N$. A consequence of the Vapnik–Chervonenkis theorem is that

$$E[\Delta(\psi_{n,C}^N, \psi_{n,C})] \leqslant 4\sqrt{\frac{\log(8eS(C,N))}{2N}}. \quad (25)$$

Combining Eqs. (22) and (25) yields

$$\theta_n - E[\theta_{n,C}^N] \leqslant \frac{\Delta(\psi_{n,C},\psi_n)}{M[\psi_0]} + 4\sqrt{\frac{\log(8eS(C,N))}{2NM[\psi_0]^2}}. \quad (26)$$

The first (constraint) summand on the right will be small if the constraint fits the signal model.

To define the shatter coefficient of a class of filters, begin by considering an arbitrary collection $\mathscr{A}$ of measurable sets on $\Re^n$. If $\{z_1,z_2,\ldots,z_k\}$ is a set of points in $\Re^n$, let

$$\eta_{\mathscr{A}}(z_1,z_2,\ldots,z_k) = \mathrm{card}\{\{z_1,z_2,\ldots,z_k\}\cap A: A \in \mathscr{A}\}, \quad (27)$$

where "card" denotes the cardinality of the class of distinct subsets of $\{z_1, z_2, \ldots, z_k\}$ created by intersection with sets in $\mathscr{A}$. The kth *shatter coefficient* of $\mathscr{A}$ is defined by

$$s(\mathscr{A},k) = \max_{z_1,z_2,\ldots,z_k} \eta_{\mathscr{A}}(z_1,z_2,\ldots,z_k). \quad (28)$$

It is possible for $s(\mathscr{A},k) = 2^k$. If $s(\mathscr{A},k) < 2^k$, then $s(\mathscr{A},j) < 2^j$ for $j > k$. The *Vapnik–Chervonenkis (VC) dimension* of $\mathscr{A}$, denoted $V_{\mathscr{A}}$, is the largest integer k for which $s(\mathscr{A},k) = 2^k$. If $s(\mathscr{A},k) = 2^k$ for all k, then $V_{\mathscr{A}} = \infty$. Shatter-coefficient bounds can be given in terms of the VC dimension [9]. For instance, for any k,

$$s(A,k) \leqslant \sum_{i=1}^{V_{\mathscr{A}}} \binom{k}{i}. \quad (29)$$

For a filter class C, define an associated class of sets by

$$\mathscr{A}_C = \{[\mathscr{K}[\psi]\cup\{0\}]\times[\mathscr{K}[\psi]^c\times\{1\}]:\psi \in C\}, \quad (30)$$

$\mathscr{K}[\psi]$ being the kernel of ψ. The shatter coefficient of C is defined by $S(C,k) = s(\mathscr{A}_C,k)$ and the VC dimension of $\mathscr{C}$ is the VC dimension of $\mathscr{A}_C$. Bounding inequalities such as the one in Eq. (29) can be used to bound the probability of Eq. (24) in terms of the VC dimension of C. Constraining the filter class constrains the VC dimension.

If there is no computationally feasible method for choosing an optimal filter from C via the empirical error estimate, then a practical algorithm might be used to approximate the empirical-error-estimate filter. There are various algorithms for approximately optimal increasing filters. When there is a large number of variables, these methods yield filters that may not minimize the empirical error estimate. The switching algorithm often yields a filter with minimal empirical error, and is usually very close, even for large windows [31]. If an algorithm yields the estimated filter $\phi_{n,C}^N$ from the sample data, M_N denotes empirical error,

$$P(M_N[\phi_{n,C}^N] \leqslant \inf_{\psi\in C} M_N[\psi] + \rho_N) \geqslant 1 - \delta_N \quad (31)$$

and $\rho_N \to 0, \delta_N \to 0$ as $N \to \infty$, then the Vapnik–Chervonenkis theorem can be modified

to state

$$P(\Delta(\phi_{n,C}^N,\psi_{n,C}) > \rho) \leqslant \delta_N + 8S(C,N)\exp[-N(\rho-\rho_N)^2/128]. \quad (32)$$

This condition means that, with probability approaching 1, the algorithm finds an operator whose empirical error exceeds the minimal empirical error over C by an amount approaching 0 [9].

While the preceding limiting results are encouraging, for very small samples they may not be helpful in choosing a constraint, especially if there is little prior knowledge as to an appropriate constraint. Returning to the general error criterion, suppose θ_n and ϕ_n are the determination coefficients corresponding to a constraint C and a stronger constraint D, meaning that $D(\boldsymbol{X}^{(n)}) \subset C(\boldsymbol{X}^{(n)})$. Then $\phi_n \leqslant \theta_n$. If ψ_n and ξ_n are the optimal filters for C and D, respectively, then $E[\Delta(\psi_n^N, \psi_n)] \geqslant E[\Delta(\xi_n^N, \xi_n)]$and the inequality is likely to be strict. It can be quite strict if D is a much stronger constraint. Thus, $E[\theta_n^N]$ can be less than $E[\phi_n^N]$. If N is small and one constraint is significantly stronger than another, then it is common for the expected sample determination coefficient for the stronger constraint to be (perhaps much) more than for the weaker constraint. Therefore it can be prudent to use strongly constrained filters. Rather than estimate θ_n by θ_n^N, since the estimators are biased low and $\phi_n \leqslant \theta_n$, a better estimator of θ_n is $\max\{\phi_n^N, \theta_n^N\}$.

6. Coefficient of determination for increasing filters

We first consider binary operators. A binary operator ψ is increasing if $\boldsymbol{x} \leqslant \boldsymbol{z}$ implies $\psi(\mathbf{x}) \leqslant \psi(\mathbf{z})$, where $\boldsymbol{x} \leqslant \boldsymbol{z}$ is defined componentwise. The *basis* of ψ is defined by $\mathscr{B}[\psi] = \mathscr{K}[\psi]^-$, where A^- is the set of minimal elements in A. ψ possesses the morphological supremum representation

$$\psi(\boldsymbol{x}) = \bigvee\{\varepsilon_{\boldsymbol{b}}(\boldsymbol{x}) : \boldsymbol{b} \in \mathscr{B}[\psi]\}, \quad (33)$$

where $\varepsilon_{\boldsymbol{b}}(\boldsymbol{x})$ denotes the *erosion* of $\boldsymbol{x}$ by the structuring element $\boldsymbol{b}$, defined by $\varepsilon_{\boldsymbol{b}}(\boldsymbol{x}) = 1$ if $\boldsymbol{b} \leqslant \boldsymbol{x}$ and $\varepsilon_{\boldsymbol{b}}(\boldsymbol{x}) = 0$ otherwise. If $\mathscr{B}[\psi] = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_r\}$, and $\boldsymbol{b}_k = (b_{k1}, b_{k2}, \ldots, b_{kn})$ for $k = 1,2,\ldots,r$, then ψ is parameterized by $\underline{\boldsymbol{b}} = (b_{11}, b_{12}, \ldots, b_{rn})$. For $0 < m < n$, nestedness of the constraint follows from the representation by letting $b_{1,m+1} = b_{1,m+2} = \cdots = b_{1,n} = b_{2,m+1} = \cdots = b_{r,n} = 0$. Indeed, the basis resulting from this transformation can be viewed as a basis for m-dimensional operators. Hence, letting $\boldsymbol{X}^{(n)}$ denote an n-component vector, $C(\boldsymbol{X}^{(m)}) \subset C(\boldsymbol{X}^{(n)})$ according to the injection $\underline{\boldsymbol{b}}^{(m)} \to \underline{\boldsymbol{b}}^{(n)}$ defined by the transformation. The constant functions $\zeta^0(\boldsymbol{x}) \equiv 0$ and $\zeta^1(\boldsymbol{x}) \equiv 1$ are increasing and hence contained in $C(\boldsymbol{X}^{(n)})$ for all n. ζ^1 and ζ^0 possess representations according to Eq. (33) by letting the basis consist of the zero vector and the basis be null, respectively.

The MAE for an increasing operator ψ with $\mathscr{B}[\psi] = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_r\}$ is given by

$$M[\psi] = \sum_{j=1}^{r} (-1)^{j+1} \times \sum_{1 \leqslant i_1 < i_2 < \cdots < i_j \leqslant r} P(\varepsilon_{\boldsymbol{b}_{i_1} \vee \boldsymbol{b}_{i_2} \vee \cdots \vee \boldsymbol{b}_{i_j}}(\boldsymbol{X}) \neq Y). \quad (34)$$

An estimate of the optimal increasing filter can be derived via a recursive formulation of the error that gives the MAE for n-observation filters in terms of $(n-1)$-observation filters [37]. It can also be derived from the unconstrained optimal filter by switching vectors in and out its kernel to arrive at the kernel of the optimal increasing filter [31]. Eq. (34) applies to the optimal increasing filter by using its basis. The coefficient of determination is thereby determined.

The general theory of mathematical morphology treats operators between lattices; [28,29,49] the theory of computational mathematical morphology concerns operators $\psi : L^n \to M$, where L and M are complete lattices [14,19,20]. Let $\psi : \Re^n \to M = \{0,1,\ldots,m\}$ be an increasing function. For $j = 1,2,\ldots,m$,ψ has *kernel sets* $\mathscr{K}_1[\psi], \mathscr{K}_2[\psi], \ldots, \mathscr{K}_m[\psi]$ defined by $\mathscr{K}_j[\psi] = \{\boldsymbol{x} \in \Re^n : \psi(\boldsymbol{x}) \geqslant j\}$. The *basis sets* are defined by $\mathscr{B}_j[\psi] = \mathscr{K}_j[\psi]^-$ for $y = 1,2,\ldots,m$. The *kernel* and *basis* of ψ are $\mathscr{K}[\psi] = \{K_1[\psi], \mathscr{K}_2[\psi], \ldots, \mathscr{K}_m[\psi]\}$ and $\mathscr{B}[\psi] = \{\mathscr{B}_1[\psi], \mathscr{B}_2[\psi], \ldots, \mathscr{B}_m[\psi]\}$, respectively. An *n-elemental erosion* is defined, for any $\boldsymbol{b} \in \Re^n$, by $\varepsilon_{\boldsymbol{b}}(\boldsymbol{x}) = 1$ if $\boldsymbol{b} \leqslant \boldsymbol{x}$ and $\varepsilon_{\boldsymbol{b}}(\boldsymbol{x}) = 0$ otherwise. As an operator, $\varepsilon_b : \Re^n \to \{0,1\}$. The basic representation theorem states: if for any $\boldsymbol{x} \in \mathscr{K}_j[\psi]$ there exists $\boldsymbol{r} \in \mathscr{B}_j[\psi]$ such that $\boldsymbol{r} \leqslant \boldsymbol{x}$, then the

increasing function $\psi : \Re^n \to M$ possesses the kernel representation

$$\psi(\boldsymbol{x}) = \sum_{j=1}^{m} \bigvee \{\varepsilon_{\boldsymbol{b}}(\boldsymbol{x}) : \boldsymbol{b} \in \mathscr{K}_j[\psi]\}. \tag{35}$$

A minimal representation results from taking the supremum over all $\boldsymbol{b} \in \mathscr{B}_j[\psi]$.

The kernel and basis representations are necessary, and their form provides a sufficient representation in the sense that, for any class of sets $\mathscr{L}_1, \mathscr{L}_2, \ldots, \mathscr{L}_m \subset \Re^n$, the operator defined by Eq. (35) with $\mathscr{L}_j$ in place of $\mathscr{K}_j[\psi]$ is increasing because $\varepsilon_{\boldsymbol{b}}$ is increasing. For the representation to be a basis representation, $\mathscr{L}_1, \mathscr{L}_2, \ldots, \mathscr{L}_m$ must satisfy two properties: (1) they must be *consistent*, meaning that, if $\boldsymbol{r} \in \mathscr{L}_j$ and $\ell \leqslant j$, then there exists $\boldsymbol{s} \in \mathscr{L}_\ell$ such that $\boldsymbol{s} \leqslant \boldsymbol{r}$; and (2) each $\mathscr{L}_j$ must be *self-minimal*, meaning $\mathscr{L}_j = \mathscr{L}_j^-$. Filter design involves finding a self-minimal, consistent class of sets that determines an estimate of the optimal increasing filter. If $\Re$ is replaced by a finite integer interval, then, as in the binary setting, there exists an extension of the error representation of Eq. (34) and a filter design algorithm based on a recursive formulation of the error [39].

To see how the Vapnik–Chervonenkis theorem applies to morphological signal processing, we apply it to increasing operators, $\psi : \Re^n \to \{0,1\}$. Eq. (35) applies with $m = 1$. For any structuring element $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)$,

$$\mathscr{K}[\varepsilon_{\boldsymbol{b}}] = \prod_{j=1}^{n} [b_j, \infty). \tag{36}$$

The VC dimension of the class of all n-products of semi-infinite intervals is n. Hence, the VC dimension of the class of all erosion kernels is n. The VC dimension of a class of complements is the same as the dimension of the class, the VC dimension is unchanged by taking a product with a singleton, and the VC dimension of a class of unions, $A_1 \cup A_2$, where $A_1 \in \mathscr{A}_1$ and $A_2 \in \mathscr{A}_2$, is bounded by the product of the VC dimensions of $\mathscr{A}_1$ and $\mathscr{A}_2$. Thus, from Eqs. (30) and (36), we conclude that the VC dimension of the class of erosions is bounded by n^2. If ψ is a supremum of M erosions, then $\mathscr{K}[\psi]$ is the union of the M kernels. Hence, the VC dimension of the class of all M-erosion kernels is bounded by n^M, and Eq. (30) shows that the VC dimension of the class of all M-erosion filters is bounded by n^{2M}.

7. Estimating the estimation error

Comparison of θ_n^N and θ_n requires estimation of $\Delta(\psi_n^N, \psi_n)$. One way is to estimate it from the sample S that gave the estimated filter ψ_n^N. This *resubstitution estimate* for $\varepsilon[\psi_n^N]$ is defined by

$$\varepsilon_S[\psi_n^N] = \frac{1}{N} \sum_{(\boldsymbol{x},y) \in S} l(y, \psi(\boldsymbol{x})). \tag{37}$$

Derivation of ψ_n^N from S is notationally implicit. The resubstitution estimate for $\Delta(\psi_n^N, \psi_n)$ is

$$\Delta_S(\psi_n^N, \psi_n) = \varepsilon_S[\psi_n^N] - \varepsilon[\psi_n]. \tag{38}$$

This leads to a corresponding resubstitution estimate, $\theta_n^{N,S}$, for θ_n.

For binary filtering with MAE, the resubstitution estimate $M_S[\psi_n^N]$ is the empirical error estimate for ψ_n^N on the data of S. It is given by Eq. (13) upon replacing all probabilities P by their estimates, P_S, from S. This yields the estimation-error estimate

$$\begin{aligned}\Delta_S(\psi_n^N, \psi_n) &= \sum_{\boldsymbol{x}} [P_S(\boldsymbol{x}) \min\{P_S(Y = 0|\boldsymbol{x}), P_S(Y = 1|\boldsymbol{x})\} \\ &\quad - P(\boldsymbol{x}) \min\{P(Y = 0|\boldsymbol{x}), P(Y = 1|\boldsymbol{x})\}].\end{aligned} \tag{39}$$

For each $\boldsymbol{x}$ observed in training, its term in the sum may be positive or negative. If N is small relative to n, then a large proportion of the vectors will not be observed in training. For these, $P_S(\boldsymbol{x}) = 0$, and the summand for $\boldsymbol{x}$ will be null. These null terms tend to make $\Delta_S(\psi_n^N, \psi_n)$ a low estimate of $\Delta(\psi_n^N, \psi_n)$. For small N, many vectors are observed only once in training. For these, $\min\{P_S(Y = 0|\boldsymbol{x}), P_S(Y = 1|\boldsymbol{x})\} = 0$. This also lowers $\Delta_S(\psi_n^N, \psi_n)$. Indeed, $\Delta_S(\psi_n^N, \psi_n)$ is typically negative, even if $\Delta(\psi_n^N, \psi_n)$ substantially exceeds 0. For an extreme case, if each $\boldsymbol{x}$ is observed at most once in training, then Eq. (39) reduces to $\Delta_S(\psi_n^N, \psi_n) = -M[\psi_n]$ and the MAE estimate is $M_S[\psi_n^N] = 0$. This situation is not rare for small N. Going further, Eq. (8) applies with θ_n^N and $\Delta(\psi_n^N, \psi_n)$ replaced by their

estimates $\theta_n^{N,S}$ and $\Delta_S(\psi_n^N, \psi_n)$. If $\Delta_S(\psi_n^N,\psi_n) < 0$, then $\theta_n^{N,S} > \theta_n$.

Relative to random sampling, θ_n^N and $\theta_n^{N,S}$ are both estimators of θ_n, and are functions of $\Delta(\psi_n^N, \psi_n)$ and $\Delta_S(\psi_n^N,\psi_n)$, respectively. For small N, θ_n^N and $\theta_n^{N,S}$ tend to be substantially low- and high-biased, respectively. The expectation $E[\Delta(\psi_n^N, \psi_n)]$ is based on the expected incremental error over the entire probability space, and is given by

$$\begin{aligned} E[\Delta(\psi_n^N,\psi_n)] &= \sum_S \left(\sum_{(x,y)} |\psi_n^N(\boldsymbol{x}) - y| P(\boldsymbol{x}, y) \right) P(S) - M[\psi_n] \\ &= \sum_S \left(\sum_x P(\boldsymbol{x})(\psi_n^N(\boldsymbol{x}) P(Y = 0|\boldsymbol{x}) \right. \\ &\quad \left. + (1 - \psi_n^N(\boldsymbol{x}))P(Y = 1|\boldsymbol{x})) \right) P(S) - M[\psi_n], \end{aligned} \tag{40}$$

where $P(S)$ is the probability of sample S, ψ_n^N depends on S, and the outer sum is over all samples. For the resubstitution estimator,

$$\begin{aligned} E[\Delta_S(\psi_n^N, \psi_n)] &= \sum_S \left(\sum_x [P_S(\boldsymbol{x}) \min\{P_S(Y = 0|\boldsymbol{x}), \right. \\ &\quad \left. P_S(Y = 1|\boldsymbol{x})\}] \right) P(S) - M[\psi_n] \\ &= \sum_S \left(\sum_x P_S(\boldsymbol{x})(\psi_n^N(\boldsymbol{x}) P_S(Y = 0|\boldsymbol{x}) \right. \\ &\quad \left. + (1 - \psi_n^N(\boldsymbol{x}))P_S(Y = 1|\boldsymbol{x})) \right) P(S) - M[\psi_n]. \end{aligned} \tag{41}$$

Hence, the difference between the expectations is

$$\begin{aligned} E[\Delta(\psi_n^N, \psi_n)] - E[\Delta_S(\psi_n^N, \psi_n)] &= \sum_S \left(\sum_x \psi_n^N(\boldsymbol{x})(P(\boldsymbol{x},0) - P_S(\boldsymbol{x},0)) \right. \\ &\quad \left. + (1 - \psi_n^N(\boldsymbol{x}))(P(\boldsymbol{x},1) - P_S(\boldsymbol{x},1)) \right) P(S). \end{aligned} \tag{42}$$

The difference results from differences between the probabilities $P(\boldsymbol{x}, y)$ and their sample estimates. The kind of reasoning applied subsequent to Eq. (39) applies to show why it is common to have a large difference when N is small relative to n. From Eq. (9), $E[\theta_n^{N,S}] - E[\theta_n^N]$ is given by the preceding difference divided by $M[\psi_0]$. From Eq. (42), $E[\theta_n^{N,S}] - E[\theta_n^N] \to 0$ as $N \to \infty$. In the small-N direction, Eq. (39) shows that, for any $\delta > 0$, there exists a largest positive integer $N_{n,\delta}$ such that $N \leqslant N_{n,\delta}$ implies $E[\Delta_S(\psi_n^N,\psi_n)] \leqslant -M[\psi_n] + \delta$. Thus, $N \leqslant N_{n,\delta}$ implies $E[\theta_n^{N,S}] \geqslant 1 - \delta$.

In general (not just for MAE), for small N we usually have $\theta_n^N \leqslant \theta_n \leqslant \theta_n^{N,S}$. The matter is clarified by expectations. $\varepsilon[\psi_n]$ can be estimated via S to obtain an estimate $\varepsilon_S[\psi_n]$. Since ψ_n^N is optimal on S, $\varepsilon_S[\psi_n^N] \leqslant \varepsilon_S[\psi_n]$. Since sampling is random, taking expectations relative to sampling yields $E[\varepsilon_S[\psi_n^N]] \leqslant \varepsilon[\psi_n]$. Therefore, owing to the optimality of ψ_n,

$$E[\varepsilon_S[\psi_n^N]] \leqslant \varepsilon[\psi_n] \leqslant E[\varepsilon[\psi_n^N]]. \tag{43}$$

Moreover, $E[\varepsilon_S[\psi_n^N]] \to \varepsilon[\psi_n]$ as $N \to \infty$, and $E[\varepsilon[\psi_n^N]] \to \varepsilon[\psi_n]$ as $N \to \infty$. Hence, the double inequality of Eq. (43) provides an envelope converging upon the error of the optimal filter [25,30,45]. Subtracting $\varepsilon[\psi_n]$ from the inequalities and applying Eq. (9) yields

$$E[\theta_n^N] \leqslant \theta_n \leqslant E[\theta_n^{N,S}]. \tag{44}$$

Moreover, $E[\theta_n^{N,S}] - E[\theta_n^N] \to 0$ as $N \to \infty$.

If there is sufficient test data, say J pairs, beyond that used for training, then we can take a standard training-testing approach by designing a filter ψ_n^N, and estimating $\varepsilon[\psi_n^N]$ and $\varepsilon[\psi_0^N]$ from the test data. An estimate $\hat{\theta}_n^{N,J}$ of θ_n is obtained from the error estimates. J has to be large enough to get good estimates of $\varepsilon[\psi_n^N]$ and $\varepsilon[\psi_0^N]$. Alternatively, we could take M random samples, $S_1, S_2, \ldots, S_M$; estimate ψ_0 and ψ_n from each S_k to obtain estimate filters $\psi_{0,1}^N, \psi_{0,2}^N, \ldots, \psi_{0,M}^N$, and $\psi_{n,1}^N, \psi_{n,2}^N, \ldots, \psi_{n,M}^N$; use independent data to obtain error estimates $\hat{\varepsilon}[\psi_{0,1}^{N,J}]$, $\hat{\varepsilon}[\psi_{0,2}^{N,J}], \ldots, \hat{\varepsilon}[\psi_{0,M}^{N,J}]$, $\hat{\varepsilon}[\psi_{n,1}^{N,J}]$, $\hat{\varepsilon}[\psi_{n,2}^{N,J}], \ldots, \hat{\varepsilon}[\psi_{n,M}^{N,J}]$; and then form the estimator

$$\hat{\theta}_n^N = \frac{1}{M} \sum_{k=1}^{M} \hat{\theta}_{n,k}^{N,J} = \frac{1}{M} \sum_{k=1}^{M} \frac{\hat{\varepsilon}[\psi_{0,k}^{N,J}] - \hat{\varepsilon}[\psi_{n,k}^{N,J}]}{\hat{\varepsilon}[\psi_{0,k}^{N,J}]}. \tag{45}$$

The principle is straightforward: if, after finding the estimate filters, we could find their exact errors

without depending on test data, then each summand in Eq. (45) would be replaced by

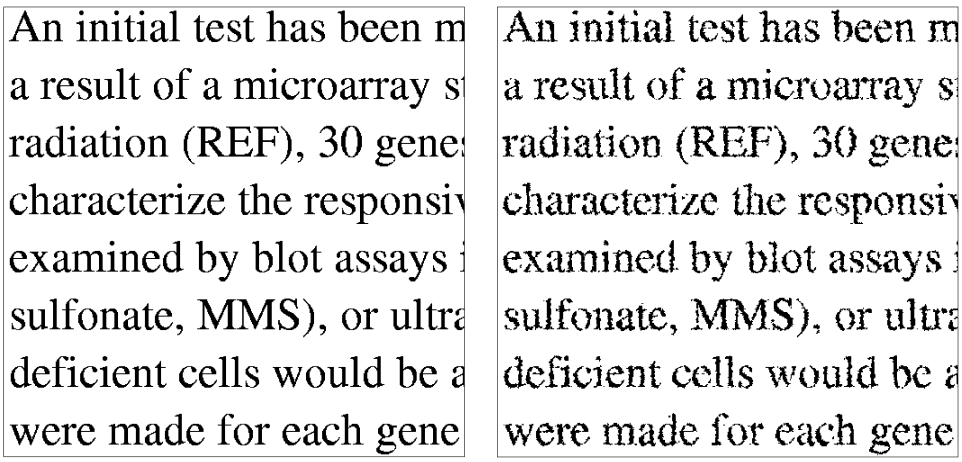
$$\hat{\theta}^N_{n,k} = \frac{\varepsilon[\psi^N_{0,k}] - \varepsilon[\psi^N_{n,k}]}{\varepsilon[\psi^N_{0,k}]}, \tag{46}$$

which is a sample value for θ^N_n. The estimator of Eq. (45) would then be the sample mean for θ^N_n, which is a consistent estimator of θ^N_n. For large test-data sets, the estimator of Eq. (45) is close to the sample mean for θ^N_n. As an estimator of θ_n, it depends on both N and J. If $N + J$ is fixed, then increasing N and decreasing J improves the designed filters while reducing the precision of the error estimates; whereas increasing J and decreasing N improves the precision of the error estimates while making the designed filters poorer estimates of the optimal filter. In either case, one effect enhances estimation of θ_n while the other diminishes it.

In practice we may have very limited data, and therefore the procedure cannot be used as stated. Instead, if there are Q sample pairs, we can proceed by sequentially randomly splitting the data into training sets $S_1, S_2, \ldots, S_M$ of size N and test sets $T_1, T_2, \ldots, T_M$ of size $J = Q - N$. Then compute Eq. (45) using the filters and errors from the training and test sets, respectively. The average provides an estimator $\tilde{\theta}^N_n$ for θ^N_n. For small N, the filters derived from the training data vary widely, as do their errors. The strategy is based on law-of-large-number effects insofar as $\tilde{\theta}^N_n$ approximates the sample mean for θ^N_n. Closeness to the sample mean depends on sufficiently large J, which is problematic. Moreover, random data splitting produces a random sample relative to the data set, not the distribution of $(\boldsymbol{X}, Y)$, and therefore the degree to which $\tilde{\theta}^N_n$ approximates the sample mean for θ^N_n depends on the degree to which the data set fits the distribution of $(\boldsymbol{X}, Y)$. The well-studied deleted-estimate approach is to let $J = 1$ and to design filters for all Q training sets having $N = Q - 1$ data pairs [9].

8. Application: image restoration

We consider estimation of the coefficient of determination for increasing window size based on differing amounts of data in the case of binary image restoration. We use the test-data estimator $\hat{\theta}^N_n(M = 64)$ and the resubstitution estimator $\theta^{N,S}_n$. Fig. 1 shows realizations of the ideal binary text process and a degraded realization formed by a well-used edge-degradation model [40]. Optimal filters have been designed using windows containing 3 through 16 pixels. In each case, the number of $(\boldsymbol{x}, y)$ pairs available is 67,320,000, and



(a) (b)

Fig. 1. (a) Ideal image and, (b) observed image (partial).

various percentages of the full set, ranging from 0.001% of the data (684 sample pairs) through 2% of the data (1,346,400 sample pairs) have been used for training. Test errors have been computed using the entire data set (which is sufficiently large that it can be considered to be the full population, meaning that $\hat{\theta}_n^N$ can be considered to be a sample mean). The 3D plot of Fig. 2 shows the computed values of $\hat{\theta}_n^N$ (lower surface) and $\theta_n^{N,S}$ (upper surface) for various window and data sizes. For fixed window size, $\hat{\theta}_n^N$ and $\theta_n^{N,S}$ increase and decrease, respectively, for increasing data size. For fixed data size, $\theta_n^{N,S}$ increases substantially for increasing window size (n); however, for $n \geqslant 5$, $\hat{\theta}_n^N$ is fairly stable.

Graphs for the fixed (symmetric) window sizes 5, 9, 13, and 17 are shown in Fig. 3. The lower and upper curves are for $\hat{\theta}_n^N$ and $\theta_n^{N,S}$, respectively, and the graphs include plots about each line showing the manner in which the 64 individual (non-averaged) estimates are dispersed for each estimator. We see that these are far more dispersed for small data sets. We also see that the estimates converge together much more rapidly for small windows, and that, for window size 17, they are still substantially separated for the largest training set. When $\hat{\theta}_n^N$ and $\theta_n^{N,S}$ have converged together, we can confidently take their common value to be θ_n. For window sizes 5, 9, 13, and 17, we have the estimates 0.77, 0.78, 0.81, and 0.82, where the latter two are guesses between the curves that have yet to converge.

The problem is with small data sets. For these, there is increasing divergence between $\hat{\theta}_n^N$ and $\theta_n^{N,S}$ as n increases. For decreasing N, we are encouraged by the stability of the test error $\hat{\theta}_n^N$, as opposed to the strongly increasing values of the resubstitution error $\theta_n^{N,S}$. There are two reasons for choosing $\hat{\theta}_n^N$ over $\theta_n^{N,S}$. It provides a better estimate of θ_n and is conservative. Nonetheless, it is biased low, and the increasing monotonicity of θ_n for increasing n can be violated by the estimates. For the present experiment, $\hat{\theta}_{13}^N = 0.75 < \hat{\theta}_5^N$ for data size $10^{-2}\%$.

9. Application: genomic control

The human genome is a highly complex nonlinear control system regulating cell functions. A primary means for regulating cellular activity is the

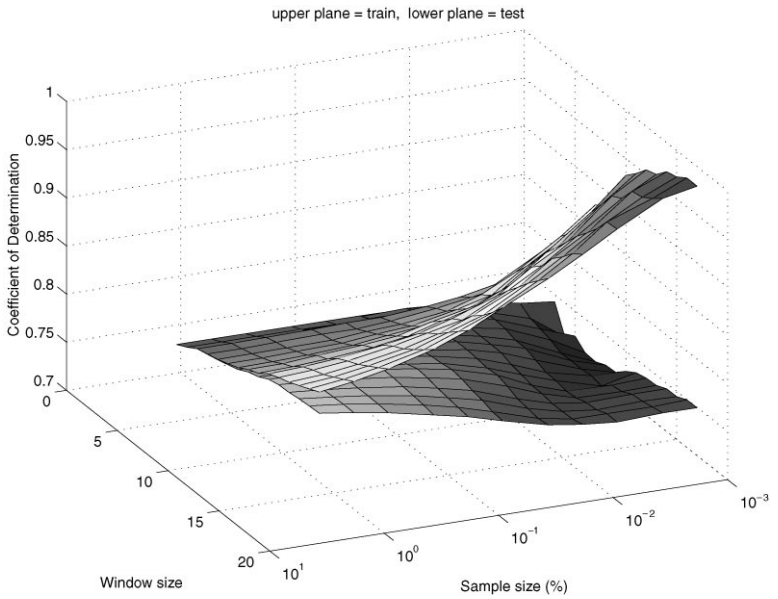


Fig. 2. Coefficients of determination for each sample and window size (3D view).

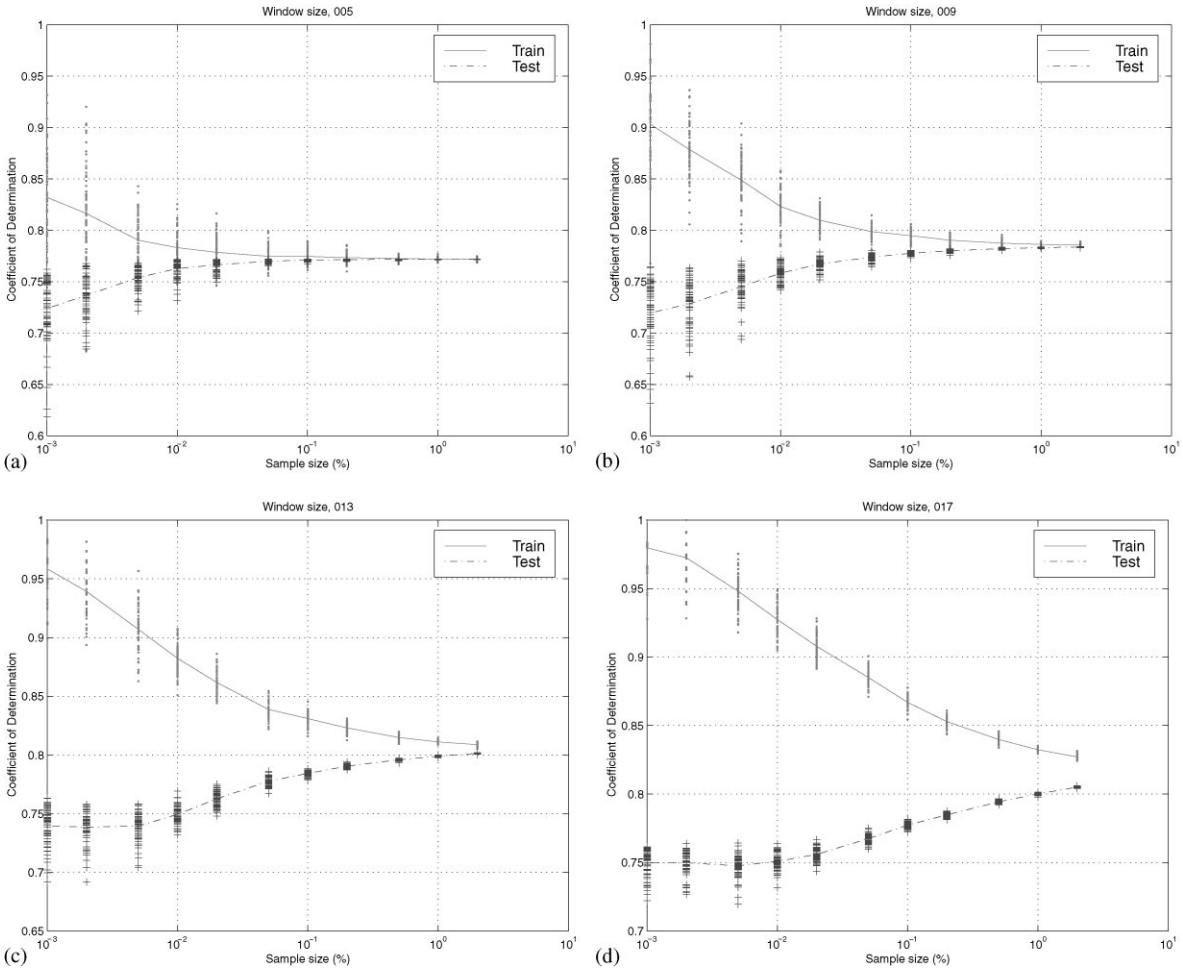


Fig. 3. Coefficients of determination for fixed window sizes: (a) 5 pixels window, (b) 9 pixels window, (c) 13 pixels window, (d) 17 pixels window.

control of protein production via the amounts of mRNA expressed by individual genes. Levels of gene expression are modulated by protein machinery that senses conditions internal and external to the cell. The tools required to build an understanding of genomic regulation of expression are those that reveal the probability characteristics of the vector random function consisting of expression levels. Basic to understanding is the ability to discover how expression levels of various genes, in conjunction with external factors, can be used to predict other expression levels. The study of expression-level prediction has recently been made possible by the development of cDNA microarrays, in which transcript levels can be determined for thousands of genes simultaneously. Microarray data has been used to design discrete nonlinear predictors (filters) whose observation variables are gene expression levels and quantifiers for external stimuli [15,32]. The data are discrete because the analog expression levels are quantized into ternary expression data: $[-1$ (down-regulated), $+1$ (up-regulated), or 0 (invariant)]. External stimuli are quantified as 1 [present] and 0 [not present]. Because there are many genes and a very small number of microarrays, it is not practical to precisely

design filters, but it is possible to estimate coefficients of determination.

We briefly describe an application to genotoxic stress analysis using the data-splitting estimator $\tilde{\theta}_n^N$. We refer to the full study for a complete description of methodology and experimental results [32]. The experiment involved twelve genes and three external conditions. Thirty microarrays were used, so that $Q = 30$. The data was randomly split into $N = 20$ training pairs and $J = 10$ test pairs. Error was MSE. $\tilde{\theta}_n^N$ was based on $M = 256$ trials. Because of the small sample size, the number of predictors was kept to $n \leqslant 4$. Both unconstrained and constrained filters were considered. The constrained filter was a ternary perceptron

$$\xi_n(\boldsymbol{X}^{(n)}) = T(a_1 X_1 + a_2 X_2 + \ldots + a_n X_n + b), \tag{47}$$

where the threshold function T is defined by $T(z) = -1$ if $z < -0.5$, $T(z) = 0$ if $-0.5 \leqslant z \leqslant 0.5$, and $T(z) = +1$ if $z > 0.5$. The optimal perceptron was estimated by a stochastic training algorithm. All possible sets of four or less predictors were used to predict all possible targets.

Owing to the small sample-data set, we expect the estimators to be quite low biased. To illustrate the effects of constraint, let θ_n and ϕ_n be the unconstrained and perceptron predictors, respectively. $\theta_n > \phi_n$, and the bias of $\tilde{\theta}_n^N$ should exceed the bias of $\tilde{\phi}_n^N$. What these biases are we have no way of determining. Even with these low biases, some fairly strong relations were observed. For predictor genes IAP-1, PC-1, and SSAT, and target gene BCL3, $\tilde{\phi}_3^N = 0.664$ and $\tilde{\theta}_3^N = 0.334$. Since $\theta_3 > \phi_3$, the error for $\tilde{\theta}_3^N$ as an estimator of θ_3 exceeds the difference, 0.330. Since a perceptron provides an approximation of a linear filter, there appears to be a somewhat linear relation between the predictors and the target. Based on the principle of taking the maximum between the estimates for a weaker and stronger constraint, we take 0.664 as our estimate for θ_3. In some cases, the inherent nonlinearity of genomic regulation can be sufficiently strong to overcome the estimation-error differential. For predictor genes RCH1, PC-1, and p53, and target gene BCL3, $\tilde{\phi}_3^N = 0.174$ and $\tilde{\theta}_3^N = 0.507$, a striking difference. For target REL-B and predictor genes BCL3 and ATF3, in conjunction with external ionizing radiation (IR), $\tilde{\phi}_3^N = 0.528$ and $\tilde{\theta}_3^N = 0.603$. There are higher estimated coefficients, for instance, $\tilde{\phi}_2^N = 0.733$ for target REL-B and predictors SSAT and IR.

10. Conclusion

Correlation can be used in linear filtering to evaluate the significance of various observation variables relative to estimating another random variable. The coefficient of determination can be used for nonlinear filtering. In areas such as image processing, the demand for larger windows continues to grow with the access to increasing computational power. Satisfactory filter design can be greatly enhanced by limiting the observations. By using the envelope convergence of training and test data estimators, it is possible to expend a great deal of simulation power on getting good estimation of the determination coefficients for various windows, and then simply restrict oneself to the most determinative windows when practical constraints require using less data. In applications such as the one from genomics, data is extremely limited and there are strong requirements for nonlinear filtering. Estimation and determination problems loom large and will no doubt receive much more attention. The determination coefficient permits biologists to focus on particular connections in the genome and coefficient estimates are useful even if they are biased and not overly precise, because at least the estimated coefficients provide a practical means of discrimination among potential predictor sets.

Because determination is based on error estimation, the matter is closely related to pattern recognition. Key differences include the kinds of constraints and design tools relevant to signal and image processing, the specifics of morphological signal representation, and the desire for certain algebraic and structural operator properties. A main focus of this paper has been to explicate some of the relevant relationships, in particular, as they apply to increasing operators and their basis/kernel representations. This is important for estimation because numerous design tools for

nonlinear operators are based on these representations.

There has also been a review of error estimation for designed filters in the context of the problems herein. Given the relatively small amounts of available sample data in applications in which the coefficient of determination may be useful, error estimation becomes an important factor. One cannot, *ipso facto*, dismiss the resubstitution estimator, because data splitting may be computationally too costly, especially when one is searching through subsets of predictor sets taken from arrays of over a thousand measurements to find determinative sets.

References

[1] J.T. Astola, P. Kuosmanen, Representation and optimization of stack filters, in: E.R. Dougherty, J.T. Astola (Eds.), Nonlinear Filters for Image Processing, SPIE and IEEE Presses, Belingham, 1999.

[2] G.J.F. Banon, J. Barrera, Minimal representation for translation-invariant set mappings by mathematical morphology, SIAM J. Appl. Math. 51 (1991) 1782–1798.

[3] G.J.F. Banon, J. Barrera, Decomposition of mappings between complete lattices by mathematical morphology, Part I. general lattices, Signal Processing 30 (1993) 299–322.

[4] J. Barrera, E.R. Dougherty, N.S.T. Hirata, Design of optimal morphological operators using prior filters, Acta Stereologica 16 (3) (1997) 183–200.

[5] J. Barrera, E.R. Dougherty, N.S. Tomita, Automatic programming of binary morphological machines by design of statistically optimal operators in the context of computational learning theory, Electronic Imaging 6 (1) (1997) 54–67.

[6] J. Barrera, N.S. Tomita, F.S. Correa da Silva, Automatic programming of morphological machines by PAC learning, Proceedings of SPIE, Vol. 2568, 1995.

[7] E.J. Coyle, J.-H. Lin, Stack filters and the mean absolute error criterion, IEEE Trans. Acoust. Speech Signal Process. 36 (8) (1988) 1244–1254.

[8] J.L. De Risi, V.R. Iyer, P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, Science 278 (1997) 680–686.

[9] L. Devroye, L. Gyorfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, New York, 1996.

[10] E.R. Dougherty, Optimal mean-square N-observation digital morphological filters – Part I: optimal binary filters, CVGIP: Image Understanding 55 (1) (1992) 36–54.

[11] E.R. Dougherty, Optimal mean-square N-observation digital morphological filters – Part II: optimal gray-scale filters, CVGIP: Image Understanding 55 (1) (1992) 55–72.

[12] E.R. Dougherty, Existence and synthesis of minimal-basis morphological solutions for a restoration-based boundary-value problem, Math. Imaging Vision 6 (1996) 315–333.

[13] E.R. Dougherty, J. Barrera, Logical binary operators, in: E.R. Dougherty, J.T. Astola (Eds.), Nonlinear filters for Image Processing, SPIE and IEEE Presses, Belingham, 1999.

[14] E.R. Dougherty, J. Barrera, Computational gray-scale operators, in: E.R. Dougherty, J.T. Astola (Eds.), Nonlinear Filters for Image Processing, SPIE and IEEE Presses, Belingham, 1999.

[15] E.R. Dougherty, M.L. Bittner, Y. Chen, S. Kim, K. Sivakumar, J. Barrera, P. Meltzer, J. Trent, Nonlinear filters in genomic control, Proceedings of the IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing, Antalya, 10–14 June, 1999.

[16] E.R. Dougherty, R.P. Loce, Optimal mean-absolute-error hit-or-miss filters: morphological representation and estimation of the binary conditional expectation, Opt. Eng. 32 (4) (1993) 815–823.

[17] E.R. Dougherty, R.P. Loce, Precision of morphological representation estimators for translation-invariant binary filters: increasing and nonincreasing, Signal Processing 40 (3) (1994) 129–154.

[18] E.R. Dougherty, R.P. Loce, Optimal binary differencing filters: design, logic complexity, precision analysis, and applications to digital document processing, Electron. Imaging 5 (1) (1996) 66–86.

[19] E.R. Dougherty, D. Sinha, Computational mathematical morphology, Signal Processing 38 (1994) 21–29.

[20] E.R. Dougherty, D. Sinha, Computational gray-scale mathematical morphology on lattices (A computer-based image algebra). Part I: architecture, Real-Time Imaging 1 (1995) 69–85.

[21] E.R. Dougherty, Y. Zhang, Y. Chen, Optimal iterative increasing binary morphological filters, Opt. Eng. 35 (12) (1996) 3495–3507.

[22] D.J. Duggan, M.L. Bittner, Y. Chen, P.S. Meltzer, J.M. Trent, Expression profiling using cDNA microarrays, Nature Gene. 21 (1999) 10–14.

[23] M. Gabbouj, E.J. Coyle, Minimum mean absolute error stack filtering with structuring constraints and goals, IEEE Trans. Acoust. Speech Signal Process. 38 (6) (1990) 955–968.

[24] C. Giardina, E.R. Dougherty, Morphological Methods in Image and Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[25] N. Glick, Sample-based multinomial classification, Biometrics 29 (1973) 241–256.

[26] N.R. Harvey, S. Marshall, The use of genetic algorithms in morphological filter design, Signal Processing: Image Communication 8 (1) (1996) 55–72.

[27] D. Hausler, Decision theoretic generalizations of the PAC model for neural nets and other learning applications, Inform. Comput. 100 (1992).

[28] H.J. Heijmans, Morphological Operators, Academic Press, New York, 1994.

[29] H.J. Heijmans, C. Ronse, I. The algebraic basis of mathematical morphology, dilations and erosions, Comput. Vision Graphics Image Process. 50 (3) (1990) 245–295.
[30] M. Hills, Allocation rules and their error rates, J. Roy. Statist. Soc. B 28 (1966) 1–31.
[31] N.S.T. Hirata, E.R. Dougherty, J. Barrera, A switching algorithm for design of optimal increasing binary filters over large windows, Pattern Recognition 33 (2000) 1052–1081.
[32] S. Kim, E.R. Dougherty, M.L. Bittner, Y. Chen, K.L. Sivakumar, P.S. Meltzer, J.M. Trent, A general framework for the analysis of multivariate gene interaction via expression arrays, Biomedical Optics (2000), in press.
[33] P. Kraft, N.R. Harvey, S. Marshall, Parallel genetic algorithms in the optimization of morphological filters: A general design tool, Electron. Imaging 6 (4) (1997) 504–516.
[34] P. Kuosmanen, J. Astola, Optimal stack filters under rank selection and structural constraints, Signal Processing 41 (3) (1995) 309–338.
[35] P. Kuosmanen, J. Astola, Breakdown points, breakdown probabilities, midpoint sensitivity curves, and optimization of stack filters, Circuits Systems Signal Process. 15 (2) (1966) 165–211.
[36] P. Kuosmanen, P. Koivisto, H. Huttunen, J. Astola, Shape preservation criteria and optimal soft morphological filtering, Math. Imaging Vision 5 (4) (1995) 319–336.
[37] R.P. Loce, E.R. Dougherty, Optimal morphological restoration: The morphological filter mean-absolute-error theorem, Visual Commun. Image Representation 3 (4) (1992) 412–432.
[38] R.P. Loce, E.R. Dougherty, Facilitation of optimal binary morphological filter design via structuring element libraries and design constraints, Opti. Eng. 31 (5) (1992) 1008–1025.
[39] R.P. Loce, E.R. Dougherty, Mean-absolute-error representation and optimization of computational-morphological filters, CVGIP: Image Understanding 57 (1) (1995) 27–32.
[40] R.P. Loce, E.R. Dougherty, Enhancement and Restoration of Digital Documents: Statistical Design of Nonlinear Algorithms, SPIE Press, Bellingham, 1997.
[41] P. Maragos, R. Schafer, Morphological filters — Part I: their set-theoretic analysis and relations to linear shift-invariant filters, IEEE Trans. Acoust. Speech Signal Process. 35 (1987) 1153–1169.
[42] P. Maragos, R. Schafer, Morphological filters — Part II: their relations to medians, order statistics, and stack filters, IEEE Trans. Acoust. Speech Signal Process. 35 (1987) 1153–1169.
[43] G. Matheron, Random Sets and Integral Geometry, Wiley, New York, 1975.
[44] P. Salembier, Structuring element adaptation for morphological filters, Visual Commun. Image Representation 3 (2) 1992.
[45] O.V. Sarca, J. Astola, On connections between robustness, precision, and storage requirements in statistical design of filters, SPIE Proceedings on Nonlinear Image Process. X, Vol. 3646 January 1999, pp. 2–13.
[46] O.V. Sarca, E.R. Dougherty, J. Astola, Secondarily Constrained Boolean Filters, Signal Processing 71 (3) (1998) 247–263.
[47] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270 (1995) 467–470.
[48] J. Serra, Image Analysis and Mathematical Morphology, Academic Press, New London, 1982.
[49] J. Serra (Ed.), Image Analysis and Mathematical Morphology, Vol. 2, Academic Press, New York, 1988.
[50] V. Vapnik, A. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. 16 (1971) 264–280.
[51] R.E. Walpole, R.H. Myers, Probability and Statistics for Engineers and Scientists, Third Edition, Macmillan, New York, 1985.
[52] L. Yin, Optimal stack filter design: A structural approach, IEEE Trans. Signal Process. 43 (4) (1995) 831–840.